Multivariate Linear Regression

Chapter 8

Multivariate Analysis

- **Every program has three major elements that might affect cost:**
  - **Size**
    - » **Weight, Volume, Quantity, etc…**
  - **Performance**
    - » **Speed, Horsepower, Power Output, etc…**
  - **Technology**
    - » **Gas turbine, Stealth, Composites, etc…**
- **So far we've tried to select cost drivers that model cost as a function of <u>one</u> of these parameters.**

$$Y_i = b_0 + b_1 X + \varepsilon_i$$

Multivariate Analysis

- What if one variable is not enough?
- What if we believe there are other significant cost drivers?
- In Multivariate Linear Regression we will be working with the following model:

$$Y_i = b_0 + b_1X_1 + b_2X_2 + \ldots + b_kX_k + \varepsilon_i$$

- What do we hope to accomplish by bringing in additional independent variables?
  - Improve ability to predict
  - Reduce variation
    - » Not total variation, SST, but rather the unexplained variation, SSE.

Multiple Regression

$$y = a + b_1x_1 + b_2x_2 + \ldots + b_kx_k + \varepsilon$$

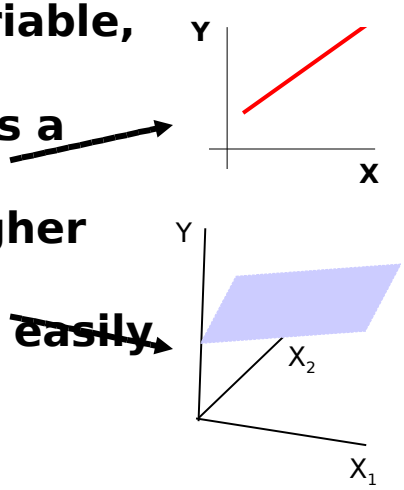
- **In general the underlying math is similar to the simple model, but matrices are used to represent the coefficients and variables**
  - **Understanding the math requires background in Linear Algebra**
  - **Demonstration is beyond the scope of the module, but can be obtained from the references**
- **Some key points to remember for multiple regression include:**
  - **Perform residual analysis between each X variable and Y**
  - **Avoid high correlation between X variables**
  - **Use the "Goodness of Fit" metrics and statistics to guide you toward a good model**

Multiple Regression

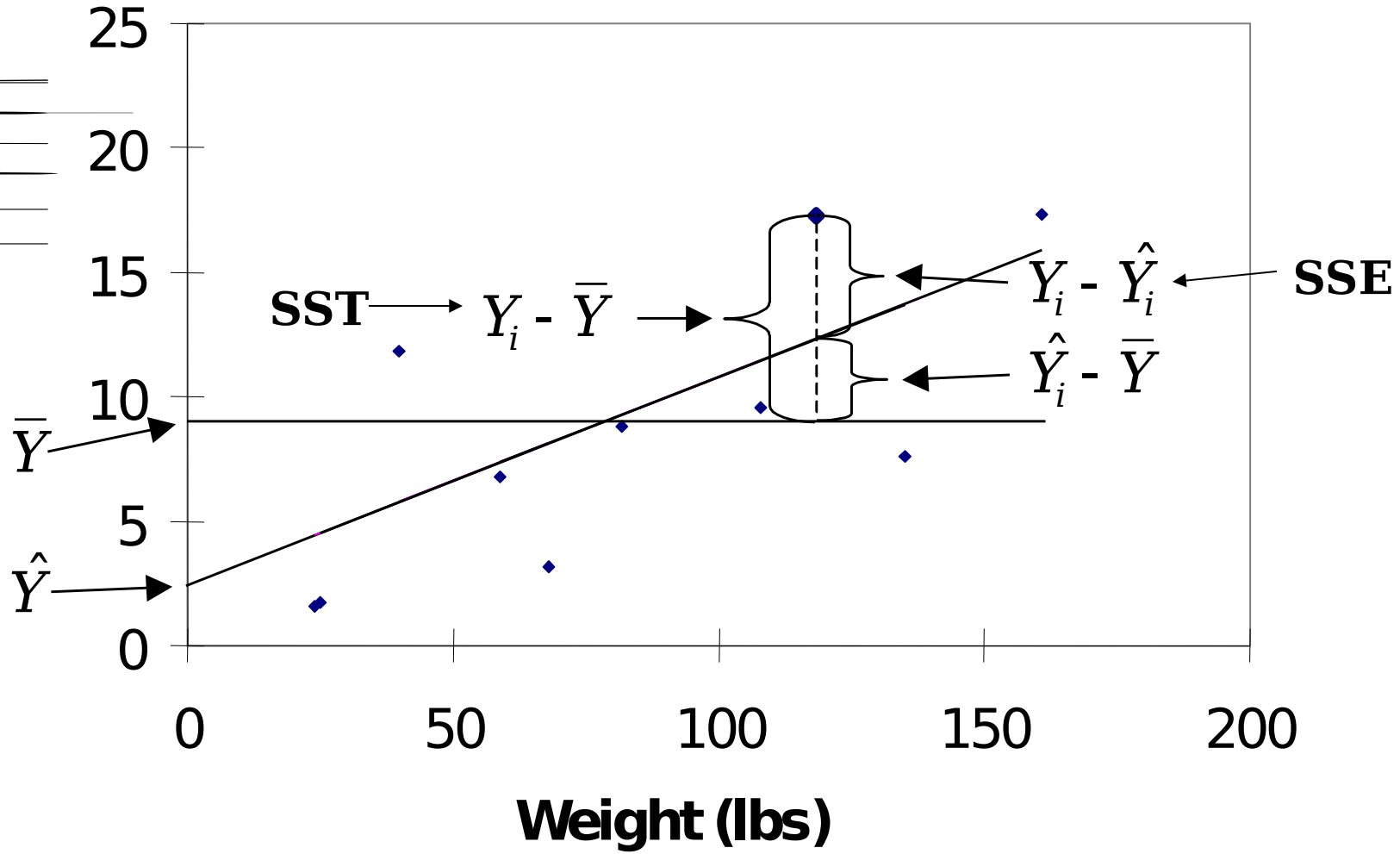
- **If there is more than one independent variable in linear regression we call it *multiple regression***
- **The general equation is as follows:**

$$y = a + b_1x_1 + b_2x_2 + \ldots + b_kx_k + \varepsilon$$

  - **So far, we have seen that for one independent variable, the equation forms a line in 2-dimensions**
  - **For two independent variables, the equation forms a plane in 3-dimensions**
  - **For three or more variables, we are working in higher dimensions and cannot picture the equation**
- **The math is more complicated, but the results can be easily obtained from a regression tool like the one in Excel**



Multivariate Analysis



Multivariate Analysis

- **Regardless of how many independent variables we bring into the model, we cannot change the total variation:**

$$SST = \sum (y_i - \bar{y})^2$$

- **We can only attempt to minimize the unexplained variation:**

$$SSE = \sum (y_i - \hat{y}_X)^2$$

- **What premium do we pay when we add a variable?**
  - **We lose one degree of freedom for each additional variable**

Multivariate Analysis

- **The same regression assumptions still apply:**
  - **Values of the independent variables are known.**
  - **The e_i are normally distributed random variables with mean equal to zero and constant variance.**
  - **The error terms are uncorrelated**

- **We will introduce Multicollinearity and talk further about the t-statistic.**

Multivariate Analysis

- **What do the coefficients, (b_1, b_2, …, b_k) represent?**
- **In a simple linear model with one X, we would say b_1 represents the change in Y given a one unit change in X.**
- **In the multivariate model, there is more of a conditional relationship.**
  - **Y is determined by the combined effects of all the X's.**
- **In the multivariate model, we say that b_1 represents the marginal change in Y given a one unit change in X_1, *while holding all the other X_i constant*.**
- **In other words, the value of b_1 is *conditional* on the presence of the other independent variables in the equation.**

Multicollinearity

- **One factor in the ability of the regression coefficient to accurately reflect the marginal contribution of an independent variable is the amount of independence between the independent variables.**
- **If X_i and X_j are statistically independent, then a change in X_i has no correlation to a change in X_j.**
- **Usually, however, there is some amount of correlation between variables.**
- **Multicollinearity occurs when X_i and X_j are related to each other.**
- **When this happens, there is an "overlap" between what X_i explains about Y and what X_j explains about Y. This makes it difficult to determine the true relationship between X_i and Y, and X_j and Y.**

Multicollinearity

- **One of the ways we can detect multicollinearity is by observing the regression coefficients.**
- **If the value of b_1 changes significantly from an equation with X_1 only to an equation with X_1 and X_2, then there is a significant amount of correlation between X_1 and X_2.**
- **A better way of detecting this is by looking at a pairwise correlation matrix.**
- **The values in the pairwise correlation matrix represent the "r" values between the variables.**
- **We will define variables as "multicollinear," or highly correlated, when $r \geq 0.7$**

Multicollinearity

- **In general, multicollinearity does not necessarily affect our ability to get a good fit, nor does it affect our ability to obtain a good prediction, *provided that we maintain the multicollinear relationship between variables*.**
- **How do we determine that relationship?**
- **Run simple linear regression between the two correlated variables.**
- **For example, if Cost = 23 + 3.5*Weight + 17*Speed and we find that weight and speed are highly correlated, then we run a regression between the variables Weight and Speed to determine their relationship.**
  - **Say, Weight = 8.3+1.2*Speed**
- **We can still use our previous CER as long as our inputs for Weight and Speed follow this relationship (approximately).**
- **If the relationship is not maintained, then we are probably estimating something different from what's in our data set.**

Effects of Multicollinearity

- **Creates variability in the regression coefficients**
  - **First, when X_1 and X_2 are highly correlated, the coefficients of each may change significantly from the one-variable models to the multivariable models.**
  - **Consider the following equations from the missile data set:**

    Cost = (-24.486) + 7.7899 * Weight

    Cost = 59.575 + 0.3096 * Range

    Cost = (-21.878) + 8.3175 * Weight + (-0.0311) * Range

  - **Notice how drastically the coefficient for range has changed.**

Effects of Multicollinearity

- **Example**

Cost	Thrust	Weight
10	7	18
20	8	44
30	17	57
30	13	67
50	22	112
60	34	112
70	39	128
80	39	165

Effects of Multicollinearity

Regression Statistics	
Multiple R	0.9781
R Square	0.9568
Adjusted R Square	0.9496
Standard Error	5.6223
Observations	8

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	4197.838	4197.838	132.799	0.000
Residual	6	189.662	31.610		
Total	7	4387.500			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	2.712	4.078	0.665	0.531	-7.268	12.691
Thrust	1.834	0.159	11.524	0.000	1.445	2.224

$$Cost = 2.712 + 1.834 \times (Thrust)$$

Effects of Multicollinearity

Regression Statistics	
Multiple R	0.9870
R Square	0.9742
Adjusted R Square	0.9699
Standard Error	4.3465
Observations	8

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	4274.147	4274.147	226.240	0.000
Residual	6	113.353	18.892		
Total	7	4387.500			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-0.4177	3.3142	-0.1260	0.9038	-8.5273	7.6920
Weight	0.5026	0.0334	15.0413	0.0000	0.4209	0.5844

$$Cost = (-0.418) + 0.503 \times (Weight)$$

Effects of Multicollinearity

Regression Statistics	
Multiple R	0.9997
R Square	0.9995
Adjusted R Square	0.9992
Standard Error	0.6916
Observations	8

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	4385.108	2192.554	4583.300	0.000
Residual	5	2.392	0.478		
Total	7	4387.500			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-0.5062	0.5274	-0.9598	0.3813	-1.8620	0.8496
Thrust	0.8291	0.0544	15.2300	0.0000	0.6892	0.9690
Weight	0.2925	0.0148	19.7856	0.0000	0.2545	0.3305

$$Cost = (-0.506) + 0.829 \times (Thrust) + 0.293 \times (Weight)$$

Effects of Multicollinearity

$$Cost = 2.712 + 1.834 \times (Thrust)$$

$$Cost = (-0.418) + 0.503 \times (Weight)$$

$$Cost = (-0.506) + 0.829 \times (Thrust) + 0.293 \times (Weight)$$

- **Notice how the coefficients have changed by using a two variable model.**
- **This is an indication that Thrust and Weight are correlated.**
- **We now regress Weight on Thrust to see what the relationship is between the two variables.**

Effects of Multicollinearity

Regression Statistics	
Multiple R	0.9331
R Square	0.8706
Adjusted R Square	0.8491
Standard Error	5.1869
Observations	8

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	1086.454	1086.454	40.383	0.001
Residual	6	161.421	26.903		
Total	7	1247.875			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	0.107	3.955	0.027	0.979	-9.571	9.784
Weight	0.253	0.040	6.355	0.001	0.156	0.351

$$Thrust \approx 0.25 \times Weigh$$

Effects of Multicollinearity

- **System 1 holds the required relationship between Weight and Thrust (approximately), while System 2 does not.**
- **Notice the variation in the cost estimates for System 2 using the three CERs.**
- **However, System 1, since Weight and Thrust follow the required relationship, is estimated fairly precisely by all three CERs.**

	System 1	System 2
Weight	95	25
Thrust	25	12
Cost (Weight)	47.33	12.15
Cost (Thrust)	48.56	24.72
Cost (Weight, Thrust)	48.01	16.76

Effects of Multicollinearity

- **When multicollinearity is present we can no longer make the statement that b_1 is the change in Y for a unit change in X_1 while holding X_2 constant.**
  - **The two variables may be related in such a way that precludes varying one while the other is held constant.**
  - **For example, perhaps the only way to increase the range of a missile is to increase the amount of the propellant, thus increasing the missile weight.**
- **One other effect is that multicollinearity might prevent a significant cost driver from entering the model during model selection.**

Remedies for Multicollinearity?

- **Drop a variable and ignore an otherwise good cost driver?**
  - **Not if we don't have to.**
- **Involve technical experts.**
  - **Determine if the model is correctly specified.**
- **Combine the variables by multiplying or dividing them.**

- **Rule of Thumb for determining if you have multicollinearity:**
  - **Widely varying coefficients**
  - **Correlation Matrix:**
    - **$r \leq 0.3$ No Problem**
    - **$0.3 \leq r \leq 0.7$ Gray Area**
    - **$r \geq 0.7$ Problems Exist**

More on the t-statistic

- **Lightweight Cruise Missile Database:**

Missile	Unit Cost (CY95$K)	Empty Weight	Max Speed	Range
A	290	39	0.7	600
B	420	54	0.66	925
C	90	16	0.84	450
D	95	15	0.59	420
E	420	57	0.37	1000
F	380	52	0.52	800
G	370	52	0.63	790
H	450	63	0.44	1600

More on the t-statistic

I. Model Form and Equation

Model Form: **Linear Model**
Number of Observations: 8
Equation in Unit Space: Cost = -29.668 + 8.342 * Weight + 9.293 * Speed + -0.03 * Range

II. Fit Measures (in Unit Space)

Coefficient Statistics Summary

Variable	Coefficient	Std Dev of Coefficient	t-statistic (coeff/sd)	Significance
Intercept	-29.668	45.699	-0.649	0.5517
Weight	8.342	0.561	14.858	0.0001
Speed	9.293	51.791	0.179	0.8666
Range	-0.03	0.028	-1.055	0.3509

Goodness of Fit Statistics

Std Error (SE)	R-Squared	R-Squared (adj)	CV (Coeff of Variation)
14.747	0.994	0.99	0.047

Analysis of Variance

Due to	Degrees of Freedom	Sum of Squares (SS)	Mean Squares (SS/DF)	F-statistic	Significance
Regression (SSR)	3	146302.033	48767.344	224.258	0
Residuals (Errors) (SSE)	4	869.842	217.46		
Total (SST)	7	147171.875			

More on the t-statistic

I. Model Form and Equation

Model Form: **Linear Model**
Number of Observations: 8
Equation in Unit Space: Cost = -21.878 + 8.318 * Weight + -0.031 * Range

II. Fit Measures (in Unit Space)

Coefficient Statistics Summary

Variable	Coefficient	Std Dev of Coefficient	t-statistic (coeff/sd)	Significance
Intercept	-21.878	12.803	-1.709	0.1481
Weight	8.318	0.49	16.991	0
Range	-0.031	0.024	-1.292	0.2528

Goodness of Fit Statistics

Std Error (SE)	R-Squared	R-Squared (adj)	CV (Coeff of Variation)
13.243	0.994	0.992	0.042

Analysis of Variance

Due to	Degrees of Freedom	Sum of Squares (SS)	Mean Squares (SS/DF)	F-statistic	Significance
Regression (SSR)	2	146295.032	73147.516	417.107	0
Residuals (Errors) (SSE)	5	876.843	175.369		
Total (SST)	7	147171.875			

Selecting the Best Model

Choosing a Model

- **We have seen what the linear model is, and explored it in depth**
- **We have looked briefly at how to generalize the approach to non-linear models**
- **You may, at this point, have several significant models from regressions**
  - **One or more linear models, with one or more significant variables**
  - **One or more non-linear models**
- **Now we will learn how to choose the “best model”**

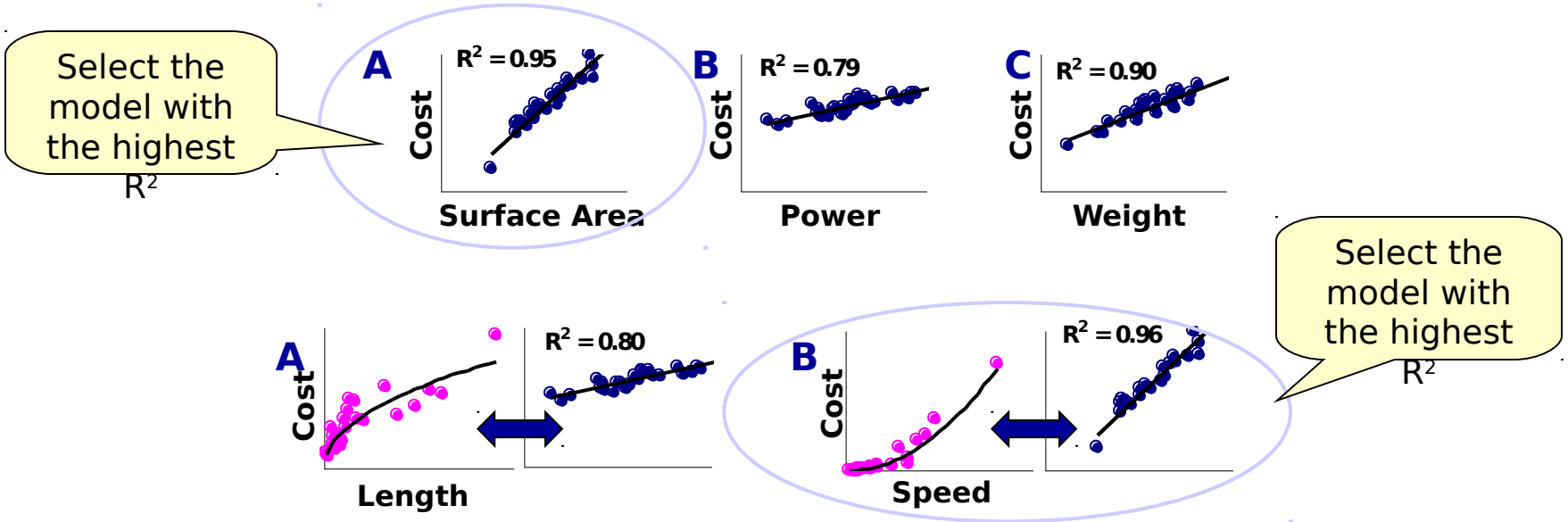
Steps for Selecting the "Best Model"

- **You should already have rejected all non-significant models first**
  - **If the F statistic is not significant**
- **You should already have stripped out all non-significant variables and made the model "minimal"**
  - **Variables with non-significant t statistics were already removed**
- **Select "within type" based on R^2**
- **Select "across type" based on SSE**

We will examine each in more detail…

Selecting “Within Type”

- **Start with only significant, “minimal” models**
- **In choosing among “models of a similar form”, R^2 is the criterion**
- **“Models of a similar form” means that you will compare**
  - **e.g., linear models with other linear models**
  - **e.g., power models with other power models**



Tip: If a model has a lower R^2, but has variables that are more useful for decision makers, retain these, and consider using them for CAIV trades and the like

Selecting “Across Type”

- **Start with only significant, “minimal” models**
- **In choosing among “models of a different form”, the SSE in unit space is the criterion**
- **“Models of a different form” means that you will compare:**
  - **e.g., linear models with non-linear models**
  - **e.g., power models with logarithmic models**
- **We must compute the SSE by:**
  - **Computing $\hat{Y}$ *in unit space* for each data point**
  - **Subtracting each $\hat{Y}$ from its corresponding actual Y value**
  - **Sum the squared values, this is the SSE**
- **An example follows…**

Warning: We cannot use R^2 to compare models of different forms because the R^2 from the regression is computed on the transformed data, and thus is distorted by the transformation